Title: Hamilton's Rule, Inclusive Fitness Maximization, and the Goal of Individual Behaviour in Symmetric Two-Player Games

Authors: Samir Okasha and Johannes Martens

Affiliation: Department of Philosophy, University of Bristol, U.K.

Running title: Hamilton's Rule and Goal of Behaviour

Corresponding author: Dr. Samir Okasha, Department of Philosophy, Cotham House, University of Bristol, Bristol BS6 6JL, U.K.

Tel: 0117 3310697; Fax: 0117 9546654; Email: Samir.Okasha@bristol.ac.uk

Abstract

Hamilton's original work on inclusive fitness theory assumed additivity of costs and benefits. Recently it has been argued that an exact version of Hamilton's rule for the spread of a pro-social allele ($rb > c$) holds under non-additive payoffs, so long as the cost and benefit terms are defined as partial regression coefficients rather than payoff parameters. This paper examines whether one of the key components of Hamilton's original theory can be preserved when the rule is generalized to the non-additive case in this way, namely that evolved organisms will behave as if trying to maximize their inclusive fitness in social encounters.

Keywords: inclusive fitness, altruism, Hamilton's rule, game theory

1 Introduction

Inclusive fitness theory is a widely-used framework for studying the evolution of social behaviour. Hamilton's original formulation of the theory contains two distinct though related ideas (Hamilton 1964, 1971). The first is Hamilton's rule, the famous criterion ($rb > c$) for when an allele coding for a social behaviour will be favoured by selection. This aspect of the theory fits with the "gene's eye" view of evolution. The second is maximization of inclusive fitness, rather than classical fitness, as the "goal" towards which an individual's social behaviour will appear designed. This aspect fits with the traditional individualist view of evolution, and is frequently employed by behavioural ecologists.

The relation between these two aspects of inclusive fitness theory is not fully settled. Much theoretical work has focused solely on the first aspect; indeed the notion of individuals "trying" to maximize their inclusive fitness is often omitted from expositions of kin selection theory. However, recently Grafen (2006, 2009), Queller (2011) and Gardner *et al.* (2011) have argued for the central importance of inclusive fitness maximization as the "goal" of individual behaviour; Grafen (2006) provides a population-genetic foundation for the idea. This goes some way to reconciling the two aspects of Hamilton's theory.

The work of Grafen, Queller and Gardner *et al.* suggests an intriguing link between social evolution and rational choice theory. For in effect, these

authors are arguing that inclusive fitness plays the role of a utility function in rational choice, i.e. it is the quantity that an evolved organism will behave as if it is trying to maximize. Thus Gardner *et al.* (2011) write: "we can imagine the individual adjusting her inclusive fitness...by altering her behaviour", before choosing an action which brings maximal inclusive fitness (p.1039-40). This way of thinking about evolution is an instance of what Sober (1988) called "the heuristic of personification", which says that a trait will be favoured by natural selection if and only if a rational individual, seeking to maximize its fitness, would choose that trait over the alternatives. In effect, Gardner *et al.* are suggesting that this heuristic is valid in social settings, where the trait in question is a social action, so long as "fitness" is defined as inclusive fitness.

Our aim here is to propose a particular way of formalizing this "rational actor heuristic" in the context of social evolution, and to ask how generally it applies. This is a pressing question because Grafen's (2006) argument that evolution will lead to inclusive fitness maximizing behaviour assumes additivity of costs and benefits. This assumption is quite restrictive since in many social situations, the benefit that a given action confers on a recipient may depend on the recipient's own type (Frank 1998, Lehmann and Rousset 2014a). In our simple model below we find that if the additivity assumption is made, then the rational actor heuristic, with inclusive fitness as the individual's utility function, applies neatly. However matters are more complex if there is non-additivity.

Asking whether the rational actor heuristic applies is different from asking whether Hamilton's rule itself applies in non-additive scenarios. This latter question has been extensively discussed in the literature. The upshot is that an exact version of Hamilton's rule does apply under non-additivity, so long as the cost and benefit terms are suitably defined (Queller 1992; Frank 1998, 2013; Gardner *et al.* 2011); though the biological significance of the resulting rule has been questioned (Allen *et al.* 2013, Birch and Okasha 2015, Birch 2015). However this does not settle the issue about individual maximization that is our focus here.

The structure of this paper is as follows. Section 2 studies social evolution using a simple additive Prisoner's dilemma, and shows how the rational actor heuristic applies to it. Section 3 considers a non-additive variant of the same model and asks whether a similar conclusion holds. Section 4 discusses the results obtained.

2 The case of additive payoffs

2.1 Additive Prisoner's dilemma

Consider a simple model of the evolution of social behaviour of the sort used in evolutionary game theory. An infinite population of haploid asexual organisms engage in pairwise social interactions in every generation. Organisms are of two types, altruists (A) and selfish (S). A types perform an action that is costly for themselves but benefits their partner; S types do not perform

the action. Type is hard-wired genetically and perfectly inherited.

An organism's payoff from the social interaction depends on its own type and its partner's type. Payoffs are interpreted as increases in lifetime reproductive fitness over a unit baseline. The social action is assumed to affect only the actor and their partner, thus local interaction is assumed absent. An A type incurs a cost of $-c$ as a result of its action and confers a benefit of b on its partner, where $c > 0$ and $b > 0$; thus the game is a Prisoner's dilemma.

Payoffs to the actor, referred to as "personal payoffs", are shown in Table 1. We let $V(i, j)$ denote the payoff to an actor from playing i when her opponent plays j, where $i, j \in \{A, S\}$. Note that payoffs are additive: an altruist alters their own payoff by $-c$ and their partners' payoff by b, irrespective of the type of their partner.

		Partner	
		A	**S**
Actor	**A**	$b - c$	$-c$
	S	b	0

Table 1: Additive Prisoner's Dilemma

There are three pair-types in the population, AA, AS and SS, whose relative frequencies in the initial generation are f_{AA}, f_{AS} and f_{SS} respectively, where $f_{AA} + f_{AS} + f_{SS} = 1$. The overall frequency of the A type in the initial generation is denoted p, where $p = f_{AA} + \frac{1}{2} f_{AS}$. The change in p over one generation is denoted Δp.

The sign and magnitude of Δp depend on the rules by which the pairs are

formed. If pairing is random, then the S type must be fitter overall, so Δp will be negative. However if pairing is assortative then the A type may be fitter overall; for the benefits of altruistic actions then fall disproportionately on other altruists. Random pairing means that the frequency distribution of the pair-types will be binomial, i.e. $f_{AA} = p^2$, $f_{AS} = 2p(1-p)$ and $f_{SS} = (1-p)^2$.

Where pairing is non-random, a simple regression analysis yields a measure of the statistical correlation between social partners. We use the variable p_i to indicate an organism's own type and p'_i to indicate its partner's type; thus $p_i = 1$ if the i^{th} organism is an A, $p_i = 0$ otherwise; and $p'_i = 1$ if the i^{th} organism is paired with an A, $p'_i = 0$ otherwise. We then compute the linear regression of p'_i on p_i, given by $b_{p'p} = Cov(p', p)/Var(p)$, which is a standard way of defining the r term of Hamilton's rule. Henceforth we refer to $b_{p'p}$ as r.

In the early kin selection literature, r was often defined in genealogical terms, e.g. as the probability that actor and partner share an allele that is identical by descent, yielding the familiar values of $\frac{1}{2}$ for full sibs, $\frac{1}{2}$ for offspring and $\frac{1}{4}$ for grandoffspring (see Michod and Hamilton 1980). In some ways this definition of r is the more natural one for expressing the idea that organisms value their relatives' reproduction in proportion to how closely related they are. However the statistical definition of r, above, yields a version of Hamilton's rule that is more generally applicable.

In the context of pairwise interactions, r can be conveniently expressed

as a difference in conditional probabilities:

$$r = Pr(\text{partner is } A \mid \text{actor is } A) - Pr(\text{partner is } A \mid \text{actor is } S)$$

It follows that r ranges from -1 (perfect disassortment) to 1 (perfect assortment); when pairing is random, $r = 0$.

2.2 Evolutionary analysis

In the Appendix, we show that the change in p over one generation is given by:

$$\Delta p = (rb - c) \cdot Var(p)/\overline{w} \tag{1}$$

where $\overline{w}$ is average population fitness. Since $Var(p)$ is non-negative, this tells us that so long as $0 < p < 1$, the A type will increase in frequency in the population whenever $rb > c$, which is of course Hamilton's rule. Since b and c are fixed parameters of the payoff matrix, this condition for the spread of A is frequency-independent so long as r itself does not change as the population evolves. Constancy of r across generations will sometimes be a reasonable assumption, for the pattern of assortment in the population, which is what r measures, may be determined by biological factors, e.g. dispersal, which are independent of the social trait that is evolving. This assumption is considered further in the Discussion section.

With the constant r assumption, the outcome of the evolutionary process is easily determined. If $rb > c$ the A type will spread to fixation; if $rb < c$ the

S type will spread to fixation; if $rb = c$ there will be no evolutionary change.

2.3 Rational actor analysis: preliminaries

To apply a rational actor analysis, we transpose our evolutionary model to a rational choice context. We consider two players playing a symmetric game. Each player has two pure strategies, A and S. If a player plays a *mixed strategy* this means that they randomize over their pure strategies; thus π_A denotes the mixed strategy in which A is played with probability π_A and S with probability $1 - \pi_A$. The payoff to a mixed strategy is then simply its expected payoff.

Each player has a utility function which measures how desirable they find the possible outcomes of the game; we assume that both players have the same utility function. Each player's goal is to maximize their utility function. One possibility is that the utility function is given by the personal payoffs in Table 1 above, in which case we write $U(i, j) = V(i, j)$, where $U(i, j)$ is the utility a player gets from playing i when their partner plays j; $i, j \in \{A, S\}$. There are other possibilities too; see below.

Once the players' utility function has been specified, the next step is to seek the Nash equilibrium (or equilibria) of the game. (A Nash equilibrium is a pair of strategies, possibly mixed, one for each player, each of which is a best response to the other.) Game theory predicts that if the players are rational, they will end up at a Nash equilibrium of the game (see for example Binmore 2007). We can then ask whether the Nash equilibria of the game

correspond to the outcomes of the evolutionary process described above. If so, we can conclude that evolution will lead organisms to behave as if trying to maximize the utility function in question.

This is a natural way of formalizing the rational actor heuristic in a game-theoretic context. It differs somewhat from Grafen's (2006) formalization of the same idea, which posits "links" between gene-frequency change and individual optimization. Our approach allows recovery of Grafen's main result; and by taking optimization to include best-response, i.e. optimal choice conditional on the other player's choice, extends easily to the non-additive case. A similar approach is found in Alger and Weibull (2012) and Lehmann *et al.* (2015).

2.3.1 Utility as inclusive fitness

One possibility to explore is that a player's utility function depends on their partner's payoff as well as their own. For example, suppose that a player's utility for any outcome is given by the quantity: personal payoff plus r times partner's payoff, i.e. $U(i,j) = V(i,j) + rV(j,i)$. Applying this transformation to the personal payoffs yields Table 2 below, which we refer to as the "inclusive fitness payoff matrix".

		Partner	
		A	**S**
Actor	**A**	$(b-c)(r+1)$	$-c+rb$
	S	$b-rc$	0

Table 2: Additive PD with inclusive fitness payoffs

This transformation was first suggested by Hamilton (1971), and has been discussed by Grafen (1979), Bergstrom (1995), Wade and Breden (1980), Day and Taylor (1998), Taylor and Nowak (2007) and Martens (2015). It is a natural formalization of the idea that an actor, in their social behaviour, will care about their partner's payoff, discounted by relatedness, as well as their own payoff. Clearly, other transformations of the personal payoff matrix are also conceivable.

The payoffs in Table 2 do not correspond exactly to the verbal definition of inclusive fitness in Hamilton (1964), which was: "the personal fitness which an individual actually expresses...once it is stripped of all components which can be considered as due to the individual's social environment...then augmented by certain fractions of the quantities of harm and benefit which the individual himself causes to the fitnesses of his neighbours...The fractions in question are simply the coefficients of relationship" (p. 8). This definition is sometimes but not always adhered in the literature.

The discrepancy between Table 2 and Hamilton's definition arises because in the left column, the actor's payoff has not been stripped of the component that is due to the partner's altruistic action (b), and has been augmented by r times the partner's entire payoff, rather than the portion of that payoff that is caused by the actor (b.) Applying Hamilton's definition exactly would lead to the payoff matrix in Table 3 below.

Note that Table 3 derives from Table 2 by subtraction of the quantity $b - rc$ from the left column. In game-theoretical terms, Table 3 is thus

		Partner	
		A	**S**
Actor	**A**	$-c + rb$	$-c + rb$
	S	0	0

Table 3: Additive PD with Hamilton (1964) payoffs

a "local shift" of Table 2 (and vice-versa), which means that their Nash equilibria are necessarily identical (Weibull 1995). Therefore, if the players' utility function is given by Table 3, game theory predicts exactly the same outcome(s) as if it were given by Table 2. So although taking Table 2 as the definition of inclusive fitness involves an element of "double counting" – which Hamilton's definition was designed to avoid – it is harmless.

In fact there is a positive reason to prefer Table 2 as the definition of inclusive fitness, in a game-theoretic context. For Hamilton's definition does not generalize easily to non-additive payoffs. With non-additivity it is unclear how to decide which component of the actor's payoff is "caused" by its partner's action and vice-versa (cf. Allen *et al.* 2013). By contrast, the definition used in Table 2 – actor payoff plus r times partner payoff – applies just as well to the non-additive case. In order not to prejudge the issue of whether inclusive fitness maximization, or a similar result, obtains under non-additivity, this is the definition preferred here.

2.4 Rational actor analysis: results

Suppose firstly that the utility function is personal payoff (Table 1). It is easy to see that (S, S) is the only Nash equilibrium of the game, since S strongly dominates A, i.e. each player does strictly better by playing S irrespective of their partner's choice. This familiar result shows that the rational actor heuristic fails for this choice of utility function, since it would have us conclude that altruism can never evolve, which we know to be false.

What if the utility function is inclusive fitness payoff (Table 2)? In that case, we can show the following. If $rb > c$ then (A, A) is the unique Nash equilibrium; if $rb < c$ then (S, S) is the unique Nash equilibrium; if $rb = c$ then (A, A) and (S, S) are both Nash equilibria, as is every pair of mixed strategies, so game theory makes no prediction about the players' choices (See Appendix for proof).

It follows that with additive payoffs, defining utility as inclusive fitness makes the rational actor heuristic valid. The condition for the A type to evolve, $rb > c$, is identical to the condition for (A, A) to be the unique Nash equilibrium of the rational game; and similarly for S (Table 4). This supports the idea that evolution will lead organisms to appear as if trying to maximize their inclusive fitness, just as Hamilton originally argued.

An equivalent perspective on the situation is this. The quantity $(rb - c)$ equals the difference in a player's inclusive fitness payoff between playing A and S, irrespective of what its partner does (see Table 2). Thus we can

$rb > c$	$\iff$	A evolves	$\iff$	(A, A) is unique Nash equilibrium
$rb < c$	$\iff$	S evolves	$\iff$	(S, S) is unique Nash equilibrium
$rb = c$	$\iff$	no evolution	$\iff$	all pairs of strategies, pure and mixed, are Nash equilbiria

Note: $\iff$ means "if and only if"

Table 4: Rational actor heuristic with utility = inclusive fitness

determine whether the A type will evolve by asking whether a rational agent, who wants to maximize their inclusive fitness, would choose A over S. In short, equating utility with inclusive fitness ensures that the rational agent's choice coincides with the "choice" made by natural selection.

2.5 A caveat: uniqueness

One important caveat is needed. In the above model, the inclusive fitness payoff matrix (Table 2) is not the unique utility function that yields the $rb > c$ condition for action A to be chosen over S. In game theory, the utility function is only ever unique up to choice of origin and unit; so any affine transformation (of the form $U' = aU + b$, where $a, b \in \mathbb{R}$, $a > 0$) will leave all Nash equilibria of the game unchanged. Furthermore, a "local shift" of the utility function, which involves adding a constant to any column of the utility matrix, will also leave unchanged the Nash equilbiria, as noted above.

One local shift of the inclusive fitness payoff matrix (Table 2) is of particular interest. If we add the quantity $(rc - rb)$ to the left-hand column of Table 2, we get the matrix in Table 5 below.

		Partner	
		A	**S**
Actor	**A**	$(b-c)$	$-c+rb$
	S	$(1-r)b$	0

Table 5: Additive PD with Grafen 1979 payoffs

The payoffs in Table 5 are related to the personal payoffs (Table 1) by the transformation $U(i,j) = rV(i,i) + (1-r)V(i,j)$. This transformation was first suggested by Grafen (1979), hence the label 'Grafen 1979 payoff'; see Bergstrom (1995), Day and Taylor (1998), Alger and Weibull (2012) for discussion. By contrast with the inclusive fitness payoffs (Table 2), which involve adding r times partner's payoff to the actor's personal payoff, the Grafen 1979 payoffs involve taking an $(r,\ 1-r)$ weighted average of the personal payoff that would accrue to the actor if their partner had chosen the same as the actor and if their partner made the choice that they actually did.

Since the Grafen 1979 payoff matrix (Table 5) is a local shift of the inclusive fitness payoff matrix (Table 2), the Nash equilibria of the resulting games are identical; thus the rational actor heuristic works equally well with either. (This is because in both cases, $(rb-c)$ is the payoff difference between playing A and S.) Therefore while our simple model has vindicated Hamilton's claim that evolution will lead organisms to behave as if trying to maximize their inclusive fitness, it is important to see that inclusive fitness (whether defined our way (Table 2) or in Hamilton's original way (Table 3)), is not the unique quantity of which this maximization claim is true.

3 Non-additive payoffs

To determine whether the above results generalize to the non-additive case, we consider a modified Prisoner's dilemma in which the payoff to an A type paired with another A type is $(b-c+d)$ rather than $(b-c)$. So the parameter d quantifies the deviation from payoff additivity, or synergistic effect, when two A types are paired together; d can be either positive or negative. The resulting payoff structure (Table 6) is sometimes referred to as a 'synergy game' (van Veelen 2009).

		Partner	
		A	**S**
Actor	**A**	$b-c+d$	$-c$
	S	b	0

Table 6: Non-additive Prisoner's dilemma ('synergy game')

Again, we assume that pairs of organisms are drawn from an infinite population to play the game; type is genetically hard-wired and mutation is absent.

3.1 Evolutionary analysis

As before, Δp denotes the change in frequency of the A type over a generation. Unsurprisingly, $rb > c$ is no longer the condition for Δp to be positive. However an exact version of Hamilton's rule can be recovered by suitably defining the cost and benefit terms, as emphasized by Gardner *et al.* (2011), whose approach we follow here. (A different approach, not discussed here,

incorporates non-additive payoffs into Hamilton's rule by a weak selection approximation; see for example Lehmann and Rousset 2014b).

For each individual i, we let w_i denote its actual reproductive fitness (number of offspring). We then write w_i as a linear regression on p_i and p'_i:

$$w_i = \alpha + b_{wp.p'} p_i + b_{wp'.p} p'_i + e_i \tag{2}$$

where α is baseline fitness; $b_{wp.p'}$ is the partial regression of an individual's fitness on their own type, controlling for their partner's type; $b_{wp'.p}$ is the partial regression of an individual's fitness on their partner's type, controlling for their own type; and e_i is the residual. These partial regression coefficients quantify the average effect (*sensu* Fisher 1930) of the actor's action, and their partner's action's, on the actor's fitness.

Following Hamilton (1964), instead of considering the effect on the actor's fitness of their partner's action $b_{wp'.p}$, we can consider the effect on their partner's fitness of the actor's action, denoted $b_{w'p.p'}$. These two partial regression coefficients are numerically identical (Taylor *et al.* 2007). (This is the well-known switch from 'neighbour-modulated' to 'inclusive' fitness.) Following Gardner *et al.* (2011), we denote the $b_{wp.p'}$ and $b_{w'p.p'}$ coefficients as $-C$ and B respectively.

Importantly, equation (2) can be fitted whether or not the true relation between w, p_i and p'_i is linear. In the non-additive case under consideration that relation is non-linear (since $d > 0$), which implies that the partial regression coefficients $-C$ and B will be functions of population-wide gene

frequencies, and liable to change as the population evolves. Therefore unlike c and b, which are fixed payoff parameters, $-C$ and B are population variables.

Following Gardner *et al.* (2007, p. 219), we can write explicit expressions for $-C$ and B in terms of r, p, and the parameters of the payoff matrix b, c and d. This yields:

$$-C = (-c) + (d) \cdot [r + p(1 - r)]/[1 + r] \tag{3}$$

$$B = (b) + (d) \cdot [r + p(1 - r)]/[1 + r] \tag{4}$$

We can then derive the following expression for evolutionary change:

$$\Delta p = (rB - C) \cdot Var(p)/\overline{w} \tag{5}$$

where $\overline{w}$ is average population fitness (see Appendix). Equation (5) tells us that when $0 < p < 1$, the A type will increase in frequency if and only if $rB > C$. This is a generalized version of Hamilton's rule, applicable whether payoffs are additive or not.

The quantity $(rB - C)$, whose sign determines whether altruism spreads, can be computed by adding equation (3) to r times equation (4). After simplifying this yields:

$$rB - C = (rb - c) + d[r + p(1 - r)] \tag{6}$$

Note that $(rB - C)$ is a function of p, so satisfaction of $rB > C$ in gen-

eration t does not imply its satisfaction in generation $t+1$. Selection is thus frequency-dependent, and neither type will necessarily spread to fixation. A polymorphic equilibrium will obtain when $p = [c - r(b+d)]/d[1-r]$; the stability of this equilibrium depends on the sign of d. The full evolutionary dynamics are summarized in Table 7 below; see Appendix for proof.

Case 1: $r < 1, d > 0$	
(i) $rb - c + rd \geq 0$	A evolves to fixation
(ii) $rb - c + d \leq 0$	S evolves to fixation
(iii) $rb - c + d > 0 > rb - c + rd$	unstable polymorphism at $p = [c - r(b+d)]/d[1-r]$
Case 2: $r < 1, d < 0$	
(i) $rb - c + d \geq 0$	A evolves to fixation
(ii) $rb - c + rd \leq 0$	S evolves to fixation
(iii) $rb - c + d < 0 < rb - c + rd$	stable polymorphism at $p = [c - r(b+d)]/d[1-r]$
Case 3: $r = 1$	
(i) $b - c + d > 0$	A evolves to fixation
(ii) $b - c + d < 0$	S evolves to fixation
(iii) $b - c + d = 0$	no evolutionary change

Table 7: Evolutionary dynamics of non-additive PD

The general version of Hamilton's rule embodied in equation (5) raises interesting interpretive questions. Some have argued that the rule in this form has little explanatory value (Nowak *et al.* 2011, Allen *et al.* 2013); while others have seen the generality of the rule as an advantage, a proof that inclusive fitness theory does not rely on restrictive assumptions (Gardner *et al.* 2011). This debate has been analyzed elsewhere (Birch 2015, Birch and Okasha 2015) and is not the focus here.

Instead our question is this. Given that equation (5) is true, and given the resulting evolutionary dynamics, can the rational actor heuristic be applied? Will evolution lead organisms to behave as if maximizing a utility function, and if so what is it?

Importantly, the answer to this question cannot simply be read off equation (5). In the additive case there was a simple link between Hamilton's rule and a utility function with the desired property: $rb-c>0$ was the condition for the A type to spread, and $(rb-c)$ the utility difference between playing A and S. One might hope to extrapolate this to the non-additive case by simply replacing $(rb-c)$ with $(rB-C)$ in Table 2. However since B and C are functions of p, they cannot meaningfully feature in the utility function.

The reason is as follows. The point of the rational actor heuristic is to find a link between gene-frequency dynamics and a "goal" that organisms behave as if they are trying to achieve. Such a link would be trivial if the "goal" were allowed to change as gene frequencies change. For the heuristic to have any value, the goal must remain fixed. So our task is to find a utility function *whose arguments are restricted to the payoff parameters* (b, c *and* d), *and the relatedness coefficient* r, which makes the rational actor heuristic work.

3.2 Rational actor analysis

To address this question, we again transpose the evolutionary model to a rational choice context and study the Nash equilibria of the resulting game.

Suppose firstly that the utility function is given by the inclusive fitness payoff transformation, i.e. personal payoff plus r times partner payoff. This yields the payoffs in Table 8 below.

		Partner	
		A	**S**
Actor	**A**	$(b-c+d)(r+1)$	$-c+rb$
	S	$b-rc$	0

Table 8: Non-additive PD with inclusive fitness payoffs

The Nash equilibria are then as follows:

(A, A) is a Nash equilibrium if and only if $rb - c + d(r+1) \geq 0$

(S, S) is a Nash equilibrium if and only if $rb - c \leq 0$

(π_A, π_A) is a mixed strategy Nash equilibrium where π_A
$= (c - rb)/d(1+r)$, so long as $0 < \pi_A < 1$.

It follows that, unlike in the additive case, the rational actor heuristic does not work when utility is defined as inclusive fitness. The condition for (A, A) to be a Nash equilibrium is not identical to the condition for A to evolve to fixation; similarly for S. Furthermore, the condition for there to be a mixed-strategy Nash equilibrium is not the same as the condition for there to be a polymorphism. So it is not true that at evolutionary equilibrium, organisms will behave as if trying to maximize their inclusive fitness.

Can we find a utility function *modulo* which the rational actor heuristic works? The answer is yes. The Grafen 1979 payoff matrix, which to recall

is derived from the personal payoff matrix by the transformation $U(i,j) = rV(i,i) + (1-r)V(i,j)$, does the trick. This yields the payoffs in Table 9 below.

		Partner	
		A	**S**
Actor	**A**	$(b-c+d)$	$-c+rb+rd$
	S	$(1-r)b$	0

Table 9: Non-additive PD with Grafen 1979 payoffs

The Nash equilibria are then as follows:

(A,A) is a Nash equilibrium if and only if $rb-c+d \geq 0$

(S,S) is a Nash equilibrium if and only if $rb-c+rd \leq 0$

(π_A, π_A) is a mixed strategy Nash equilibrium, where π_A $= (c-r(b+d))/d(1-r)$, so long as $0 < \pi_A < 1$.

This restores the rational actor heuristic. In particular, if (A,A) is the only pure-strategy Nash equilibrium, then A evolves to fixation; if (S,S) is the only pure-strategy equilibrium, then S evolves to fixation. If there is a mixed-strategy Nash equilibrium but no pure strategy equilibria, the population evolves to a stable polymorphism; if there is a mixed-strategy Nash equilibrium and both (A,A) and (S,S) are pure-strategy equilibria, then there is an unstable polymorphism; in both cases, the weights on A and S in the mixed-strategy Nash equilibrium equal the proportions of A and S in the polymorphism. Thus there is a tight correspondence between the

Nash equilibria and the evolutionary dynamics, summarized in Table 10 (see Appendix for proof)

(A, A) is only pure N.E.	$\Longrightarrow$	A evolves to fixation
(S, S) is only pure N.E.	$\Longrightarrow$	S evolves to fixation
(π_A, π_A) is only N.E.	$\Longleftrightarrow$	stable polymorphism at $p = \pi_A$
(π_A, π_A), (A, A), (S, S) all N.E.	$\Longleftrightarrow$	unstable polymorphism at $p = \pi_A$

Note: $\pi_A = (c - r(b + d)/d(1 - r))$

Table 10: Rational actor heuristic, utility = Grafen 1979 payoff

The upshot is that with non-additive payoffs, the rational actor heuristic will work so long as the utility function is defined as Grafen 1979 payoff, rather than inclusive fitness payoff. Again any affine transformation of the Grafen 1979 payoff matrix, or any local shift, will also preserve the correspondences above. Note that, unlike in the additive case, the Grafen 1979 payoff matrix (Table 9) is *not* a local shift of the inclusive fitness payoff matrix (Table 8). This is why the rational actor heuristic fails if utility is defined as inclusive fitness in the non-additive case.

4 Discussion

Hamilton's original formulation of inclusive fitness theory assumed additivity of costs and benefits. A number of authors have emphasized that an exact version of Hamilton's rule holds with non-additive payoffs, so long as the $-C$ and B terms are suitably defined. Here we have focused on the relevance of payoff additivity not for Hamilton's rule itself, but for Hamilton's (logically

distinct) claim that evolution will lead organisms to behave as if trying to maximize their inclusive fitness, understood here to mean personal payoff plus r times partner payoff.

In a recent critique, Allen *et al.* (2013) observe that arguments for inclusive fitness maximization all rely on payoff additivity, and that where selection is frequency-dependent, fitness maximization need not generally occur. They write: "evolution does not, in general, lead to the maximization of inclusive fitness or any other quantity" (p. 20138).

Our analysis partly supports this conclusion. Here we have understood maximization to include best-response, so that the presence of frequency-dependence does not automatically preclude a maximization principle from holding; and we have allowed the utility function to be any function of the payoff parameters b, c and d and the relatedness coefficient r. At the evolutionary equilibrium of our simple non-additive model, it is not true that organisms behave as if trying to maximize their inclusive fitness payoff. However there is a somewhat similar quantity – Grafen 1979 payoff – that organisms do behave as if they are trying to maximize.

It is an open question whether our positive result – maximization of Grafen 1979 payoff – extends to more complicated models of social evolution, e.g. that incorporate local interaction, multiple social partners, or class structure, or to more realistic genetic architectures than haploid inheritance. There is no guarantee that it does, as such models typically lead to more complicated evolutionary dynamics than those assumed here. As has been

emphasized before, a valid maximization argument must always deduce the quantity being maximized, if any, from the underlying evolutionary dynamics (Mylius and Diekmann 1995).

Also, we have assumed that the coefficient of relatedness, r, remains constant as the population evolves. Without this assumption, it makes little sense to allow the utility function to depend on r, as this would be tantamount to positing a changing "goal" so would again trivialize the rational actor heuristic. In some inclusive fitness models, r is in fact a dynamic variable rather than a constant (e.g. van Baalen and Rand 1998), so it cannot be assumed that our results, or ones like them, can be derived for these models.

Our negative result, that maximization of inclusive fitness only holds with additive payoffs, is in line with previous results by Bergstrom (1995) and Lehmann and Rousset (2014a); it supports some of the claims made by opponents of inclusive fitness theory such as Allen and Nowak (2015). The key logical point to note is that although a version of Hamilton's rule is indeed a fully general evolutionary principle, as Gardner *et al.* (2011) stress, no principle about individual maximization can be deduced directly from this form of the rule. Whether such a principle holds, and if so what the quantity being maximized is, needs to be shown on a case-by-case basis.

Finally, what are the implications for biological practice? Behavioural ecologists have often used inclusive fitness maximization as a way to interpret observed behaviour in the field, in line with Hamilton's original suggestion. Our analysis suggests that this will not always be possible. If an observed

social behaviour fails to maximize an individual's inclusive fitness, defined as personal payoff plus r times partner's payoff, the behaviour may nonetheless be adaptive and the population at an evolutionary equilibrium. Moreover the quantity we have called "Grafen 1979 payoff" will serve the needs of the behavioural ecologist seeking to identify the "goal" of evolved behaviour in a broader range of cases than will inclusive fitness itself.

Acknowledgements

Thanks to Ken Binmore, Bengt Autzen, Jonathan Birch, Steve Frank, Herb Gintis, Alan Grafen, Andy Gardner and two anonymous referees for their comments and discussion. This work was supported by the European Research Council Seventh Framework Program (FP7/20072013), ERC Grant agreement no. 295449.

References

Alger, I. & Weibull, J. W. 2012. A generalization of Hamilton's rule–love others how much? *J. Theor. Biol.* **299**: 42-54.

Allen B., & Nowak M.A. 2015. Games among relatives revisited. *J. Theor. Biol.* **194**: 391-407.

Allen B., Nowak M.A., & Wilson, E.O. 2013. Limitations of inclusive fitness. *Proc. Natl. Acad. Sci. USA* **110**: 21035-20139.

Bergstrom, T. 1995. On the evolution of altruistic ethical rules for siblings. *Am. Econ. Rev.* **85**: 58-81.

Binmore, K. 2007. *Playing for Real.* Oxford University Press, Oxford.

Birch J. 2015. Hamilton's rule and its discontents *Brit. J. Philos. Sci.*, in press.

Birch J. & Okasha, S. 2015. Kin selection and its critics. *Bioscience* **65**: 22-32.

Day, T. & Taylor, P.D. 1998. Unifying genetic and game theoretic models of kin selection for continuous traits. *J. Theor. Biol.* **194**: 391-407.

Fisher R.A. 1930. *The Genetical Theory of Natural Selection.* Clarendon Press, Oxford.

Frank S.A. 1998. *Foundations of Social Evolution.* Princeton University Press, Princeton, New Jersey.

Frank S.A. 2013. Natural selection VII. History and interpretation of kin selection theory. *J. Evol. Biol.* **26**: 1151-1184.

Gardner A., West S.A., & Barton, N. 2007. The relation between multilocus

484 population genetics and social evolution theory. *Am. Nat.* **169**: 207-226.

Gardner A., West S.A., & Wild, G. 2011. The genetical theory of kin selection.
486 *J. Evol. Biol.* **24**: 1020-1043.

Grafen A. 1979. The hawk-dove game played between relatives. *Anim. Behav.*
488 **27**: 905-907.

Grafen A. 2006. Optimization of inclusive fitness. *J. Theor. Biol.* **238**: 541-563.

490 Grafen A. 2009. Formalizing Darwinism and inclusive fitness theory. *Phil. Trans. R. Soc. B*: **364**: 3135-3141.

492 Hamilton W.D. 1964. The genetical evolution of social behaviour. *J. Theor. Biol.* **7**: 1-52.

494 Hamilton W.D. 1971. Selection of selfish and altruistic behaviour in some extreme models. In: *Narrow Roads of Gene Land Volume 1*, pp. 198-228. W. H.
496 Freeman, New York.

Lehmann L., Alger, I. and Weibull, J. 2015. Does evolution lead to maximizing
498 behaviour. *Evolution* **69**: 1858-1873.

Lehmann L., & Rousset, F. 2014a. Fitness, inclusive fitness and optimization.
500 *Biol. Philos.* **29**: 181-195.

Lehmann L., & Rousset, F. 2014b. The genetical theory of social behaviour.
502 *Phil. Trans. R. Soc. B* **369**: 20130357.

Martens, J. 2015. Hamilton meets causal decision theory. *Brit. J. Philos. Sci.*, in
504 press.

Michod, R. & Hamilton, W.D. 1980. Coefficients of relationship in sociobiology.

Nature **288**: 694-697.

Mylius, S.D., & Diekmann, O. 1995. On evolutionarily stable life histories, optimization and the need to be specific about density dependence. *Oikos* **74**: 218-224.

Nowak, M.A. & Tarnita, C.E. & Wilson, E.O. 2011. Nowak et. al. reply. *Nature* **471**: E9-E10.

Price, G.R. 1970. Selection and covariance. *Nature* **227**: 520-521.

Queller D.C. 1992. A general model for kin selection. *Evolution* **46**: 376-380.

Queller D.C. 2011. Expanded social fitness and Hamilton's rule for kin, kith and kind. *Proc. Natl. Acad. Sci. USA* **108**: 10792-10799.

Sober, E. 1988. Three differences between evolution and deliberation. In: *Modeling Rationality, Morality and Evolution* (P. Danielson, ed.), pp. 408-422. Oxford University Press, Oxford.

Taylor C., & Nowak, M.A. 2007. Transforming the dilemma. *Evolution* **61**: 2281-2292.

Taylor P.D., Wild, G. & Gardner, A. 2007. Direct fitness or inclusive fitness: how shall we model kin selection? *J. Evol. Biol.* **20**: 301-309.

van Baalen, M. & Rand, D.A. 2007. The unit of selection in viscous populations and the evolution of altruism. *J. Theor. Biol.* **193**: 631-648.

van Veelen M. 2009. Group selection, kin selection, altruism and cooperation: when inclusive fitness is right and when it can be wrong. *J. Theor. Biol.* **259**: 589-600.

528 Wade, M., & Breden, F. 1980. The evolution of cheating and selfish behaviour. *Behav. Ecol. Sociobiol.* **7**: 167-72.

530 Weibull J. 1995. *Evolutionary Game Theory.* MIT Press, Massachusetts.